Target-aware Lattice Rescoring for Dialect Recognition

Rong Tong[1,2], Bin Ma[1], Haizhou Li[1,2] and Eng Siong Chng[2]

[1] Institute for Infocomm Research, Singapore 138632
[2] School of Computer Engineering, Nanyang Technological University, Singapore
{tongrong,mabin,hli}@i2r.a-star.edu.sg, aseschng@ntu.edu.sg

Abstract

We observed that human listeners distinguish one dialect from another by paying special attention to some particular phonetic and/or phonotactic patterns. Motivated by this observation, we propose a technique that emulates this process. We explore a target-aware lattice rescoring (TALR) process that revises the *n*-gram statistics in a lattice with target dialect information. We then derive *n*-gram statistics as the phonotactic features from the lattice and develop a system under the vector space modeling framework. The experiment results show that the proposed technique consistently improves dialect recognition performance on 30-second test utterances. We achieved equal error rates (EERs) of 4.57% and 13.28% with 3-gram statistics for Chinese and English dialect recognition in 2007 NIST Language Recognition Evaluation 30-second closed test sets.

Index Terms: speech recognition, dialect recognition, spoken language recognition, lattice rescore, language model

1. Introduction

The task of dialect recognition is to confirm which variation of a language is spoken in a speech utterance. Two terms are commonly used to describe the variations within a language: accent and dialect. In general, accent refers to the phonetic variation of words in the same language, while dialect refers to the variation of pronunciation, grammar and vocabulary over different versions of a language. In this paper, we do not make an explicit distinction between dialect and accent. We use *dialect* to refer to both types of variations. Dialect recognition is a more challenging task than the general language recognition as the difference among dialects is usually more subtle than that among languages.

Studies on dialect recognition were greatly inspired by the advances in language recognition [1-3]. Many techniques have been developed for dialect recognition. Recently, people have started looking into how to equip acoustic or phonotactic models with target dialect information. This is to emulate a human listening process where people pay special attention to some phonetic and phonotactic patterns when differentiating languages or dialects. Shen et al. [4] proposed a dialect recognition system that is based on dialect specific acoustic models. The approach starts with a language specific phone recognizer. Speech data for target languages are transcribed into phone sequences which are used for unsupervised adaptation of phone models. This technique incorporates target language information into the acoustic modeling of phone recognizer. Biadsy et al. [5] also studied ways to adapt acoustic models with target dialect information over four Arabic dialects.

In phonotactic feature based language recognition, usually we don't apply language model because a phone recognizer in the front-end is supposed to process speech in multiple languages [1,4,5]. In our previous work, with a set of target languages in mind, we proposed to derive parallel phone tokenizers with target-aware language models (TALM) from an existing phone recognizer [9]. This technique allows us to optimize a PPR-VSM system [8] for a set of target languages.

A phone lattice is a graph that presents the possible phone paths among the feature frames. It is reported that phonotactic information derived from lattice achieves better performance than that from 1-best phone sequence [6]. Siniscalchi et. al [7] proposed to rescore the lattice with additional knowledge to improve the speech recognition accuracy. Inspired by the successes of lattice rescoring in speech recognition and language recognition, we extend the idea of TALM by incorporating linguistic information in lattice generation for dialect recognition.

In a phone recognition front-end, we generate a phone lattice with a null grammar, i.e., a free phone loop grammar. Hence the generated lattice is an acoustic lattice as no language model is involved. In this paper, we study a way to derive multiple dialect-dependent lattices from a single acoustic lattice. Given a phone recognizer, we train a language model using the decoding results of held out data from a target dialect, the language model is then used to rescore the original acoustic lattice to obtain dialect-dependent phonotactic statistics. We call the proposed technique target-aware lattice rescoring (TALR). It offers several advantages: 1) TALR derives phonotactic statistics that are relevant to the target dialects of interest, that allow us to optimize the systems for target dialects; 2) TALR derives multiple phonotactic statistics from a single phone tokenizer through a rescoring process, which is computationally efficient. 3) No additional transcribed data is required for TALR to work.

This paper is structured as follows. In Section 2, we describe the mathematical formulation of dialect recognition with phonotactic features. In Section 3, we present the proposed target-aware lattice rescoring approach for dialect recognition. We introduce the experimental setup and report the dialect recognition experimental results on three sets of dialects in Section 4. Finally we conclude in Section 5.

2. Dialect Recognition with Phonotactic Features

Without loss of generality, we consider a single phone recognizer as tokenization front-end for dialect recognition. The phonotactic features are obtained in two steps: i) an input speech sample is firstly decoded by the phone recognizer; ii) the phonotactic features presented by *n*-gram phone statistics are then derived from the decoding results.

Given a speech segment O, the objective of a dialect recognition system is to find the dialect d^* as follows,

$$d^* = \arg\max_d \sum_S p(O \mid \mathrm{H}, S) p(S \mid d) \quad (1)$$

where H denotes a set of phone models, d denotes the target dialect, S is the phone segmentation obtained from the phone recognizer. It is a common practice to use a free phone loop grammar in the decoding process, hence no language related information is involved when decoding S. As the summation over all possible segmentation S is not practical in real applications, recently studies in language recognition offer two solutions for this.

1) Equation (1) is approximated by using just the 1-best phone decoding results as follows,

$$d^* \cong \arg\max_d p(O \mid \mathrm{H}, S^*) p(S^* \mid d) \quad (2)$$

Here S^* denotes the best phone sequence obtained from the phone recognizer. We derive the n-gram statistics from the phone sequence [1,10].

2) Equation (1) is approximated by using n-gram statistics derived from phone lattice [6], an intermediate result during phone decoding. Specifically, we compute the expectation of all phone segmentation in the lattice:

$$d^* \cong \arg\max_d \sum_W E_\Omega[p(W \mid H)] p(W \mid d) \quad (3)$$

where E_Ω denotes the expectation in lattice Ω and the sum is performed on all the possible n-grams W in the lattice. It is known that the phonotactic features from lattice offer a better performance than those from 1-best phone sequence in language recognition [6]. It is also generally agreed that the accuracy of phone recognition and the quality of the lattices are critical to language recognition [4,6].

3. Rescoring Lattice for Dialect Recognition

One of the important issues in phonotactic dialect recognition is how to identify the most discriminative features that are good in differentiating one dialect from others. In human listening experience, we observed that human listeners distinguish one language or dialect from another by paying special attention to particular phonetic and phonotactic patterns. Motivated by this observation, we propose techniques that emulate this process.

3.1. Target-aware language models

In our previous work, we study a target-aware language models (TALM), that was designed to pay special attention to target languages for language recognition [9]. TALM learns the discriminative statistics across languages and incorporates them into the decoding process in the phone recognition frontend. With TALM, the speech utterances are first processed by a phone recognizer to generate phone sequences. We then measure the discriminative ability of each phone as to how they separate a target language from others. In this way, we generate a unigram language model for each target language, and obtain a new set of phone recognizers which share the same acoustic model, but differentiated by language models.

3.2. Target-aware lattice rescoring

Inspired by TALM, we now bring the idea of target-awareness for dialect recognition. We rescore the language-independent lattice using dialect-dependent language models. By rescoring a lattice, we revise the posterior probability of phone n-gram in the lattice. Equation (3) is therefore re-written as:

$$d^* \cong \arg\max_d \sum_W E_\Omega[p(W \mid H, L)] p(W \mid d) \quad (4)$$

where L denotes the language model used in the phone decoding process.

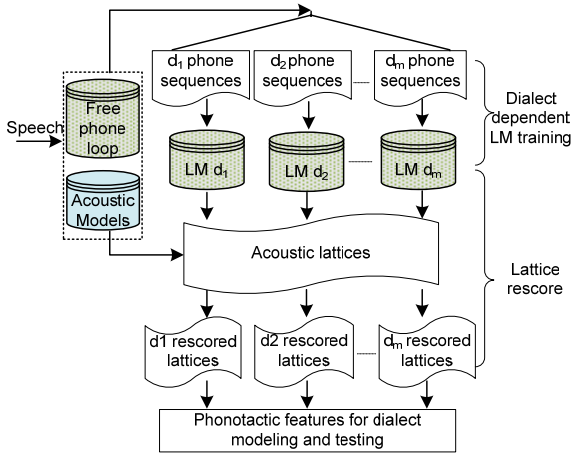

Figure 1 TALR: a target-aware process that rescores a phone lattice with different dialect-dependent language models.

Figure 1 depicts the target aware lattice rescoring (TALR) process. The rescoring process during the system development involves the following 4 steps. The dialect dependent language models are trained through the first two steps, and the lattice rescoring is performed on training and testing data with step 3 and 4.

1) Decode a held out set of dialect data using the phone recognizer with a free phone loop grammar;

2) Train a n-gram language model for each of the target dialects with the phone sequences derived from step 1. For m target dialects, we obtain m dialect-dependent language models: $\{LM_{d1}, LM_{d2}, ..., LM_{dm}\}$;

3) Decode the dialect data using phone recognizer, an acoustic lattice is generated for each utterance;

4) Rescore the acoustic lattices with the dialect dependent language model $\{LM_{d1}, LM_{d2}, ..., LM_{dm}\}$ obtained in step 2.

In this way, we obtain a collection of dialect-dependent lattices from each of the dialect-dependent language models, which are used for development of PPR-VSM dialect recognition system [8].

While sharing a similar idea with TALM, TALR is fundamentally different in the following two aspects: First, we construct higher order n-gram language model in TALR to capture detailed phonotactic information, as opposed to unigram language model in TALM. Second, we apply different n-gram language models to the same lattice during rescoring in TALR, while we apply different n-gram language models in TALM in different phone sequence decoding processing. In other words, TALR only requires one-time acoustic decoding, while TALM requires multiple acoustic decoding. As the lattice rescoring is computationally efficient, TALR involves little additional computation at run-time.

3.3. Language modeling for dialects

We propose a 2-step language modeling for dialects. In the first step, we train a general language model by including all dialects of a language in the training process. In the second step, we adapt the language model with dialect specific data

through a MAP process [12]. This is to benefit from the a more robust general language model and to overcome the problem of insufficient dialect training data.

Grammar scale factor [14] is a parameter in speech recognition that regulates the contributions of acoustic and language model. In TALR, we use the grammar scale factor to regulate the contribution of the dialect-dependent language models during the lattice rescoring, with respect to the acoustic models. In the lattice, the likelihood of an arc is described as:

$$\log(p(O \mid H)) + \log(p(O \mid L)) * \beta \quad (5)$$

where β denotes the grammar scale factor, which is studied in the dialect recognition experiments.

4. Experiments

4.1. Experiment setup

We use the PPR-VSM [8] system architecture in all the experiments. The BUT Hungarian phone recognizer [10] is used as the front-end. Unlike other conventional phone recognizers, the acoustic features in the BUT phone recognizers are extracted over a long temporal context. Three neural networks (NN) are trained to produce the phoneme posterior probabilities. Based on our experiment results on same data set, the Hungarian phone recognizer achieved better performance than the 6 HMM based phone recognizers combined system in language recognition [9,15]. In the lattice generation process, the Hungarian phone recognizer first tokenizes a speech utterance into posteriors, the HTK tool [14] is then used to convert estimated posteriors into lattice, and to rescore the lattice. The SRI lattice tool kit [11] is further used to derive *n*-gram counts from the lattice.

We conduct experiments on three dialect recognition tasks in 2007 NIST Language Recognition Evaluation (LRE07). The first task is Chinese dialects recognition. It includes four dialects: Mandarin, Cantonese, Min and Wu. In this test, both Mainland and Taiwan Mandarin are categorized as Mandarin. The second task is Mandarin recognition in which only two dialects: Mainland Mandarin and Taiwan Mandarin are involved. The third task is English dialect recognition which includes two dialects of English: American English and Indian English. The number of trials for each of three sets of dialect recognition tasks is shown in Table 1.

Table 1. Dialect recognition tasks

Task	Dialects	Number of trials
Chinese (CH)	Cantonese	80
	Mandarin	158
	Min	80
	Wu	80
Mandarin (MA)	Mainland	80
	Taiwan	78
English (EN)	American	80
	Indian	160

The training data includes the English and Chinese portions of LDC CallFriend corpus, the story data of OGI 22-language corpus [13], OHSU 2005[1] and LRE 07 development data sets released by LDC[2]. The dialect recognition performance is reported in terms of equal error rate (EER) and DET curves. In all the experiments, we assume that the priors for target and non-target dialects are equal.

The dialect language model training data for Mandarin and American English are randomly selected from CallFriend corpus. 400 segments from each of the dialect data are held out for dialect-dependent language model training. Each of the 400 segments contains about 30 seconds of speech. For other dialects: Indian English, Cantonese, Min and Wu, the language model training data are selected from OHSU and LRE 07 development set. As the amount of data are limited, only 100 segments for each dialect are selected.

4.2. Experiment Results

4.2.1 TALR vs. baseline system

To evaluate the effective of TALR in dialect recognition, we compare the dialect recognition performance with a baseline system. The baseline system is a PPR-VSM system with Hungarian phone recognizer as the tokenization front end. A free loop phone grammar is used in the decoding process. The phonotactic information are derived from the lattice. In the lattice rescoring process, the grammar scale factor β is set to 15 for all the three tasks, and the 3-gram dialect-dependent language models are used in lattice rescoring.

Table 2. EER(%) of the proposed TALR and the baseline system on three dialect recognition tasks

System/ EER(%)	30-second	10-second	3-second
Baseline CH	5.02	11.96	23.19
TALR CH	4.57 (8.96%)	11.48 (4.01%)	23.66 (-2.02%)
Baseline MA	37.34	38.61	44.94
TALR MA	22.79 (38.97%)	28.48 (26.24%)	34.81 (22.54%)
Baseline EN	14.38	25.71	30.34
TALR EN	13.28 (6.95%)	24.53 (4.59%)	29.22 (3.69%)

Table 2 compares the dialect recognition results of TALR and the baseline system on three dialect recognition tasks. The dialect recognition results on three different lengths of the test utterances are reported. The numbers in the bracket indicate the relative improvements of TALR over the baseline system.

It is noted that the dialect recognition performance are improved across most of the conditions, with the only exception of the Chinese dialect 3-second task. The EER improves the most for 30-second test conditions and the improvement decreases as the length of the test utterances becomes shorter. This may be due to the fact that the language model is more effective over a longer context.

Another observation from the Table 2 is that the Mandarin dialect recognition obtained more improvements than Chinese dialect recognition. This can be interpreted as that the Mandarin dialects have more training data for dialect-dependent language models, while 3 out of 4 dialects in Chinese dialect recognition have much less language model training data.

4.2.2 N-gram language modeling

In this section, we study the effects of different language models for dialect recognition. We apply language models

[1] http://www.ohsu.edu/.
[2] http://www.nist.gov/speech/tests/lre/2007/.

with different levels of complexity on the three dialect recognition tasks. We also compare the dialect language modeling using two steps method as described in Section 3.3 and the direct language modeling method without adaptation.

Table 3 reports the EERs of the three dialect recognition tasks with the lattice rescoring using different language models on 30-second test sets. The last column of the Table 3 shows the results of using 3-gram language model trained directly from dialect phone sequences, without any adaptation. The middle three columns show the dialect recognition results with dialect language models adapted from the dialect-independent language model.

Table 3 EER(%) with different dialect language models on 30s tests

Task/ LM	3-gram w/MAP	2-gram w/MAP	1-gram w/MAP	3-gram wo/MAP
CH	4.57	5.25	5.70	4.73
MA	22.79	22.79	24.06	23.03
EN	13.28	13.75	14.75	14.31

The experiment results show that higher order *n*-gram models achieve better performance, while moving from 2-gram to 3-gram language models makes little difference. Comparing the two 3-gram language model results, one can note that the adaptation process works favorably across all the three test sets. It confirms the need of having reliable phonotactic statistics through a robust general language model.

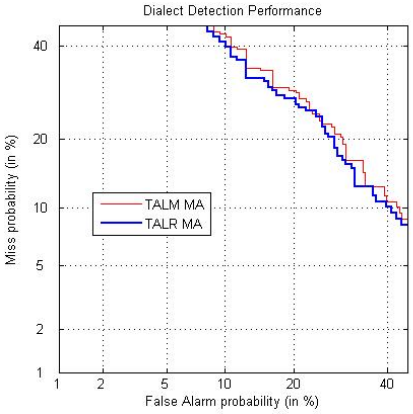

Figure 2 DET curves of TALR and TALM methods for the 30s test set of LRE07 Mandarin dialect recognition

4.2.3 TALR vs TALM

In this experiment, we further compare TALR and TALM for dialect recognition. Figure 2 illustrates the DET curves of the two methods on Mandarin dialect recognition of LRE07 30-second test set. Both methods use same sets of training and development data, and the phonotactic information is derived from lattices. TALM uses unigram language models which models the discriminative ability of phones, while TALR method uses 3-gram language models trained from the decoding results. We note that two methods achieves comparable performance on the Mandarin dialect recognition, with TALR giving slightly better performance. This should be attributed to the higher order *n*-gram statistics in lattice rescoring process. We have similar observations for other dialect groups in the LRE07 test set.

5. Conclusions

This paper studies the way to incorporate dialect information into lattice for dialect recognition. The experiment results show the accuracy of the dialect recognition is improved by rescoring lattice with dialect-dependent language models. A dialect language model is trained from the decoding results, no additional language model training data is required. As the rescoring is done on an acoustic lattice, the same lattice can be used for rescoring of many different dialects. Therefore, TALR incurs minimum additional computation. The experiment results show that a general language model is beneficial for the development of dialect language models. As expected, higher order language models lead to better dialect recognition performance.

The target-aware lattice rescoring technique is proved to be effective in dialect recognition. It achieves comparable performance with TALM method which focuses on discriminative ability of phones. As TALR doesn't require multiple acoustic decoding for the same phone recognizer, TALR is computationally efficient than TALM. As a the future work, we would like to explore the way to incorporate discriminative information in dialect dependent-language model training.

6. References

[1] M. A. Zissman, T. Gleason, D. Rekart, and B. Losiewicz, "Automatic dialect identification of extemporaneous conversational, Latin American Spanish speech", *ICASSP 1996*, Atlanta, USA

[2] P.A. Torres-Carrasquillo, T. P. Gleason, and D. A. Reynolds, "Dialect identification using Gaussian Mixture Models," in proceeding of *the speaker and language recognition workshop*, Spain, 2004

[3] J. Hou, Y. Liu, F. Zheng, J. Olsen and J. Tian, "Using Cepstral and prosodic features for Chinese accent identification", *ISCSLP 2010*, pp.177-181, December, 2010, Taiwan

[4] W. Shen, N. Chen, and D. Reynolds, "Dialect recognition using adapted phonetic models," in Proceedings of *INTERSPEECH*, Brisbane, Australia, 2008

[5] F. Biadsy, H. Soltau, L. Mangu, J. Navratil, J. Hirschberg, "Discriminative phonotactics for dialect recognition using context-dependent phone classifiers", *Odyssey 2010*, pp.263-270, June 2010

[6] J. L. Gauvain, A. Messaoudi and H. Schwenk, "Language recognition using phone lattices", *ICSLP 2004,* pp. 1283-1285, 2004

[7] S.M. Siniscalchi, C-H. Lee, "A study on integrating acoustic-phonetic information into lattice rescoring for automatic speech recognition", *Speech communication*, Vol. 51, issue 11, Nov 2009

[8] H. Li, B. Ma, and C.-H. Lee, "A Vector Space Modeling Approach to Spoken Language Identification", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 15, No. 1, Jan 2007

[9] R. Tong, B. Ma, H. Li and E. S. Chng, " Target-Aware Language Models for Spoken Language Recognition", *Interspeech 2009*, pp. 200-203, Sep. 2009

[10] P. Matejka, P. Schwarz, J. Cernocky, P. Chytil, "Phonotactic Language identification using high quality phoneme recognition", *Interspeech 2005,* pp. 2237-2240, 2005

[11] A. Stolcke, "SRILM - An Extensible Language Modeling Tooklit", In *ICSLP, 2002*, pp. 901--904, 2002

[12] M. Bacchiani and B. Roark. "Unsupervised language model adaptation", In Proceedings of *ICASSP 2003*, pp. 224–227

[13] T. Lander and R. Cole and B. Oshika and M. Noel, "The OGI 22 language telephone speech corpus", *Eurospeech 1995*, pp. 895-898, 1995

[14] http://htk.eng.cam.ac.uk

[15] R.Tong, B. Ma, H. Li and E.S. Chng, "Selecting Phonotactic Features for Language Recognition",Interspeech 2010, Sep. 2010, Japan.